HHS Public Access

Author manuscript

IEEE Trans Biomed Eng. Author manuscript; available in PMC 2018 May 01.

Published in final edited form as:

IEEE Trans Biomed Eng. 2017 May ; 64(5): 1089–1098. doi:10.1109/TBME.2016.2590950.

Automated Assessment of Disease Progression in Acute Myeloid Leukemia by Probabilistic Analysis of Flow Cytometry Data

Bartek Rajwa,
Bindley Bioscience Center, Purdue University, W. Lafayette, IN, 47907 USA

Paul K. Wallace,
Department of Flow and Image Cytometry, Roswell Park Cancer Institute, Buffalo, NY, USA

Elizabeth A. Griffiths, and
Department of Medicine, Roswell Park Cancer Institute, Buffalo, NY, USA

Murat Dundar [Member, IEEE]
Department of Computer and Information Sciences Indiana University - Purdue University, Indianapolis, IN, 46202 USA

Abstract

Objective—Flow cytometry (FC) is a widely acknowledged technology in diagnosis of acute myeloid leukemia (AML) and has been indispensable in determining progression of the disease. Although FC plays a key role as a post-therapy prognosticator and evaluator of therapeutic efficacy, the manual analysis of cytometry data is a barrier to optimization of reproducibility and objectivity. This study investigates the utility of our recently introduced non-parametric Bayesian framework in accurately predicting the direction of change in disease progression in AML patients using FC data.

Methods—The highly flexible non-parametric Bayesian model based on the infinite mixture of infinite Gaussian mixtures is used for jointly modeling data from multiple FC samples to automatically identify functionally distinct cell populations and their local realizations. Phenotype vectors are obtained by characterizing each sample by the proportions of recovered cell populations, which are in turn used to predict the direction of change in disease progression for each patient.

Results—We used 200 diseased and non-diseased immunophenotypic panels for training and tested the system with 36 additional AML cases collected at multiple time points. The proposed framework identified the change in direction of disease progression with accuracies of 90% (9 out of 10) for relapsing cases and 100% (26 out of 26) for the remaining cases.

Conclusions—We believe that these promising results are an important first step towards the development of automated predictive systems for disease monitoring and continuous response evaluation.

Significance—Automated measurement and monitoring of therapeutic response is critical not only for objective evaluation of disease status prognosis but also for timely assessment of treatment strategies.

Index Terms

minimal residual disease; flow cytometry; acute myeloid leukemia; AML; nonparametric Bayesian; Dirichlet process

I. Introduction

Acute myeloid leukemia (AML) is a malignant disease affecting both children and adults, with an age-adjusted incidence of 3.51 per 100,000 men and women per year in the United States (from 1975 to 2011). Although the 5-year relative survival rate has significantly improved since the 1970s, only 26% of patients diagnosed with AML will survive 5 years after diagnosis. For elderly patients (65+) the 5-year survival rate remains less than 6% [1].

Flow cytometry (FC) is a leading technology for cell analysis, allowing rapid evaluation of heterogeneous cellular populations in a single-cell setting; i.e., FC separately interrogates every individual cell. The analysis process uses fluorescently labeled antibodies to tag cellular epitopes known from their association with a specific cell lineage, function, or state. In combination with various probes for cell viability, structure, and function this methodology can provide information-rich data sets describing the phenotypic effects of various natural physiological phenomena or the impact of external perturbants on characteristics of cell populations [2].

Following the endorsement of the Bethesda International Consensus Group, as well as the guidance for classification of hematolymphoid neoplasms introduced by the 4th edition of the The World Health Organization's Classification of Tumours of Haematopoietic and Lymphoid Tissues, FC immunophenotyping became a standard tool for AML diagnosis and disease monitoring [3], [4], [5]. FC is universally recognized as the method of choice for determining the blast lineage as well as for detecting aberrant antigenic phenotypic profiles [6], [7]. Although molecular classification has become a new standard for disease evaluation in AML, FC immunophenotypic characterization remains crucial for early staging of the disease, monitoring response to therapy, detecting minimal residual disease (MRD), and tracking relapse or progression [4], [8]. FC evaluation of AML has been demonstrated to be highly correlated with event-free survival (time to induction failure, relapse, secondary malignancy, or death), in contrast to the only very limited prognostic value offered by morphologic studies or PCR results [9], [10].

Evaluation of response to chemotherapy and assessment of disease progression are essential tasks in the management of patients with AML. With the advent of personalized medicine and the increasing use of targeted therapies whose activity depends upon the sustained expression of surface markers identified by FC, the use of clinical-decision support systems implemented as software support in co-developed companion diagnostic devices will be absolutely necessary [11], [12].

Although FC has been demonstrated by a number of reports to be more predictive and consistent than the alternatives, practitioners acknowledge that the reported results of FC-based diagnostics depend strongly on correct interpretation of the highly complex raw FC

results. Indeed, the current approach to FC data analysis involves a painstaking interactive process of manually building long chains of gates by drawing them on the screen of a computer terminal with the help of a pointing device. The quality of the process depends immensely on the judgment and experience of the operator. Additionally, the process is hard to generalize, standardize, and transfer between laboratories or even between FC instruments designed by different manufacturers. Although the biological reasoning behind gating strategies is well established, the praxis of manual analysis remains difficult, time consuming, highly operator-dependent, costly, and hard to implement in new environments.

Ever since automated data analysis, machine learning, and bioinformatics began to establish themselves in the field of cytometry [13], [14], [15], the FC community has been trying to address these issues by proposing a number of diverse semiautomated algorithms that first pre-process the data by spectral unmixing (compensation) and then mimic the gating process with the help of a number of clustering techniques [15], [16], [17], [18], [19]. Although an increasing number of published computer-science methods leave no doubt that significant progress has been made in incorporating these techniques into analytical research pipelines, computational cytometry so far has had little or no impact on the clinical field. We hypothesize two principal reasons for that. First, even though the proposed algorithms automate the feature-extraction process, translating the raw cytometry data into vectors of cell-type proportions, the logic of these processes (the algorithm design) is ultimately driven by the experience of the operators (domain knowledge) and by their expectations regarding the data, rather than by the data alone. Therefore, the use of automated FC analysis often simply shifts the point of operator input within the analysis pipeline, rather than utilizing purely data-driven approaches. Second, the implemented machine-learning (ML) techniques typically frame the problem as a binary classification in which the patients are considered either “healthy” or “sick.” In fact, that was the framework of the FlowCAP competition, in which members of the computational cytometry community tested their solutions for automated AML classification [15]. However, as we argued before, such a setting is far from realistic. AML comprises a heterogeneous group of clonal neoplasms that differ substantially in cause, age of onset, clinical features, and prognosis. The changes during the progression of the disease, the presence of MRD, or the appearance of previously unobserved leukemia-associated phenotypes (LAPs) make FC classification of AML a more complex machine-learning problem.

Therefore, because of the tremendous diversity in the disease-associated phenotypes in AML, the computer-assisted approaches that perform well with one data set may not demonstrate such success in a clinical setting owing to the bias of the classifiers. We have already shown that a novel solution using a one-class classifier paired with non-parametric Bayesian models incorporating random effects can surpass traditional ML techniques when exposed to more challenging scenarios [20].

Herein, we report a solution to a complementary problem in automated classification - limited access to “normal” phenotypes. The traditional supervised methodology for 2-class classification operates under an assumption that both classes are well defined and that there is a certain level of homogeneity of the features within a class. As mentioned before, the diversity of AML renders this assumption invalid in relation to the characteristics of the

diseased class. We argued before that this problem can be addressed using a framework in which the diseased state is defined as a departure from normality [20].

Computational biologists often face an opposite dilemma. Although over a long period of time the researchers may successfully acquire a sufficiently diverse data set of abnormal cases representing well the complex landscape of a studied disease, access to “normal” samples may remain limited owing to prohibitive costs of sample collection, inconvenience, or ethical boundaries of research on healthy subjects. Often this problem is solved in an ad hoc manner by reusing samples representing unrelated diseases as approximately “normal.” In certain areas the problem of a small number of training samples can also be addressed in silico by using tools such as data augmentation.

In this report we also employ these strategies of using related biological data sets to define a “non-diseased” phenotype, in order to predict the direction of change and patient prognosis. The first classifier uses peripheral blood samples, which are labeled using a panel identical to that used for AML bone-marrow samples. The second classifier uses bone marrow data from patients diagnosed with lymphoma. Although the multidimensional point cloud defined by these samples does not represent a “normal” phenotype of bone marrow, it provides a direct contrast to an abnormal AML phenotype. Therefore, these cohorts of samples can be designated as “non-diseased” or “non-AML” for the purpose of classification.

The pooled diverse-diseased and non-diseased data are processed jointly using our previously demonstrated Bayesian modeling algorithm, ASPIRE [20]. The overall philosophy of this technique fundamentally differs from that of the methodology used in both traditional manual and most automated FC algorithms. First, ASPIRE does not employ any domain knowledge-based assumptions regarding the presence or frequency of particular leukemia-associated phenotypes (LAPs). We are aware that the diversity of the disease and the inherent noise in data covering a period of several years can make any approaches based on a carefully drafted set of steps in a decision tree-like arrangement impractical and highly dependent on the available training data. Instead, ASPIRE deduces the multidimensional arrangement of the data point cloud forming clusters and meta-clusters in the feature space. This is achieved by employing a non-parametric Bayesian (NPB) model as a core method in ASPIRE. The main building block of the algorithm is the highly flexible infinite mixture of infinite Gaussian mixtures (I^2GMM) [21]. Unlike traditional mixture modeling, which requires information about the possible number of clusters (relevant biological populations, in this case) in advance, ASPIRE predicts the most probable number of relevant cell populations in a heterogeneous sample while simultaneously performing model inference and assessing the proportion of cells in each of these populations. Once the distributions of cell populations are estimated on the training data set by ASPIRE, individual cell data on both training and testing cases are classified using the recovered distributions to obtain phenotype vectors characterizing data for each patient at each time point. Then, training phenotype vectors are used for training a logistic regression classifier which is validated on the test set by analyzing the change in probability values corresponding to phenotype vectors of the same patient obtained at different time points. Data collection and methods are presented in Section II. Results are discussed in Section III.

II. Materials and Methods

A. Data Collection

The study was approved by the Institutional Review Board of the Roswell Park Cancer Institute (IRB# BDR0517). The initial training set of healthy-donor blood samples (n=100) and patient bone-marrow samples (n=100) obtained at time of diagnosis were collected in sodium heparin and processed by the flow cytometry laboratory within 24 hours. The second set of model-validation samples were obtained from patients at time of diagnosis, post-induction 1 (mean day 48.5; range 26–84) and post-induction 2 (approximately 1 year post induction, mean day 293.6; range 76–731). The post-induction 2 time point is available only for the relapse cases.

Blood and bone marrow were stained, acquired, and prepared for flow cytometric analysis as previously described [22]. Briefly, bone-marrow samples were first passed through a 35-micron sieve to exclude spicules; then both blood and bone marrow were washed once with PBS containing heparin (10 units/mL) and then twice with FCM buffer (0.5% BSA, 0.004% Na_2EDTA and 0.1% sodium azide prepared in PBS, pH 7.2). Cells were resuspended to their original volume in FCM buffer and incubated with normal mouse IgG (10μg/test) for 10 min to block FC receptors. Next, washed cells were aliquoted into tubes and stained with saturating amounts of the 4-color monoclonal antibody (mAb) cocktails described in Table 1. The cells were incubated in the dark at ambient temperature for 20 minutes. After this incubation, red blood cells were lysed with BD FACSLyse (BD Bioscience, San Jose, CA) washed once with FCM buffer, and then fixed in 0.5% methanol-free formaldehyde (Polysciences, Warrington, PA). Samples were stored in the dark at 4 °C no longer than 24 hours until analysis. Cytofluorometric analysis was performed using an unmodified FACSCalibur (BD BioSciences) flow cytometer and conventional data analysis was performed using WinList (Verity Software House, Topsham, ME).

B. FC Measurement Data Structure

Each patient sample or panel obtained at a given time point was subdivided into 7 tubes and analyzed with different mAb combinations using 4 mAbs per tube. Marker combinations for each tube are shown in Table I. In addition to four fluorescence-intensity values, the measurement also provides small-angle and large-angle light-scatter characteristics. Therefore, FC data obtained for each tube contained information for 20,000 – 50,000 cells, each represented by six parameters. Data obtained from each subject at a given time point contained seven data matrices of similar sizes, one for each tube, to be analyzed for phenotypic characterization.

C. Phenotypic Characterization for Training Cases

For identifying functionally distinct cell types across all FC data matrices in the training cohort (diseased and non-diseased samples) ASPIRE algorithm is used [20]. ASPIRE assumes each FC data matrix is generated by a global latent mixture model with potentially infinitely many components. The data in each FC matrix are generated from a specific subset of components in this mixture model. A component can also be understood as a global cluster each representing potentially functionally distinct cell types; local clusters are their

realizations in individual FC measurements. In other words, a metacluster represents a generalization of a specific cellular phenotype, whereas a cluster is a group of cells in a particular sample belonging to this immunophenotype.

The Dirichlet-process mixture model (DPM) is the building block of our data model. A DPM is a mixture model with a Dirichlet-process (DP) prior defined over mixture components [23]. DPM belongs to a group of non-parametric Bayesian models. It is non-parametric because the number of clusters can arbitrarily grow to better accommodate data as needed. One of the two parameters of a DP prior, also known as the concentration parameter, controls the prior probability of producing a new component and thus indirectly controls the total number of components produced. The second parameter – the base distribution – defines the Bayesian aspect of DPM. In the case of Gaussian components one can utilize the base distribution to encode the existing knowledge of the domain by defining prior distributions over the mean vectors and covariance matrices of components.

In ASPIRE each FC data matrix is modeled by the infinite mixture of Dirichlet-process Gaussian-mixture models [21]. Model learning, which is performed in a single unified process, involves three main tasks: recovering local DPMs, finding global cluster associations of DPMs, and identifying the total number of clusters and their proportions in each FC measurement. The infinite mixture of DPMs, also known as the infinite mixture of infinitely many Gaussian mixture models (I^2GMM), which is considered as a two-layer Gaussian mixture with an arbitrarily large number of components in each layer, offers extreme flexibility in modeling data sets with skewed and multi-mode cluster distributions. The lower layer estimates the density of the overall data set by clustering individual data points to components, while the upper layer associates components with clusters to allow for cluster recovery. As local distributions of a cell type in each FC measurement are noisy realizations of the true class distribution, i.e., a global meta-cluster, we introduce a sharing mechanism to create dependencies across DPMs associated with the same cell type. This is achieved by centering the base distributions of DPMs associated with the same cell type across FC measurements on a unique global parameter, which itself is distributed according to a higher-level DPM. This global DPM not only associates local distributions of a global cluster with one another but also models the number and proportions of clusters in each FC data matrix. We use a collapsed Gibbs sampler to perform inference [20].

This two-layer non-parametric model of Gaussian mixtures is ideally suited for identifying cell types in a FC data cohort for the following reasons. As a non-parametric model it addresses sample variability resulting in multi-mode and skewed cluster distributions by employing mixture models with arbitrarily large number of components. As a hierarchical model it allows for information sharing within as well as between samples, facilitating the discovery of rare populations. As a Bayesian approach, through use of prior distributions over cluster parameters, it makes possible the direct incorporation of domain knowledge about certain cell types into the model.

The generative model that encompasses these intricacies is described next.

$$\begin{aligned} \boldsymbol{x}_{jki} &\sim p(\cdot|\theta_{jki}) \\ \theta_{jki} &\sim G_{jk} \\ G_{jk} &\sim DP(F_{\phi_k}, \alpha) \\ \phi_k &\sim G_0 \\ G_0 &\sim DP(H, \gamma) \end{aligned} \quad (1)$$

We denote each data point i of cluster k in FC data matrix j by $\boldsymbol{x}_{jki} \in \Re^d$, where $i = \{1,\ldots, n_{jk}\}$, $k = \{1, \ldots, K\}$, and $j = \{1, \ldots, J\}$, n_{jk} is the number of points from cluster k in data matrix j, K is the total number of global clusters identified, and J is the total number of data matrices. We use $\boldsymbol{\phi}_k$ and θ_{jki} to define parameters of global clusters and their local realizations in each data matrix, respectively. The parameters α and γ adjust the prior probability of creating a local and global clusters, respectively.

Both G_0 and G_{jk} are random probability measures distributed independently and identically according to a Dirichlet process. They can be considered as a mixture distribution with infinitely many components with their parameters drawn from the base distribution of the Dirichlet process and their weights from a stick-breaking distribution [24]. The stick-breaking distribution considers a unit-length stick that is broken according to a sample drawn from a Beta distribution. Unlike continuous distributions, the probability of drawing the sample twice from G_0 and G_{jk} is not zero and is proportional to the weight of the corresponding component. Thus, both G_0 and G_{jk} are considered discrete distributions and offer clustering properties.

The base distribution H of the global Dirichlet process (DP) prior is defined as follows.

$$p(\boldsymbol{\mu}, \sum) = N\left(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \frac{\sum}{\kappa_0}\right) \times W^{-1}\left(\sum|\sum_0, m\right) \quad (2)$$

where $\boldsymbol{\mu}_0$ is the expected center of the global clusters and κ_0 is a scaling constant that controls their dispersion with respect to this center. The parameter σ_0 is a positive definite matrix that encodes our prior belief about the shape of the clusters. The parameter m is a scalar that is negatively correlated with the degrees of freedom. In other words, the larger the m, the less σ will deviate from the expected shape, and vice versa.

Individual DPMs associated with the same global cluster share the same $\boldsymbol{\phi}_k$ across data matrices. The notation $F_{\boldsymbol{\phi}_k}$ indicates a distribution F centered at $\boldsymbol{\phi}_k$ and defines global cluster–specific base distributions of individual DPMs. Although $F_{\boldsymbol{\phi}_k}$ is same for all DPMs associated with the same global cluster, local clusters in data matrices are generated i.i.d. given $\boldsymbol{\phi}_k$ of corresponding DPMs. Thus, each local realization of a global cluster is modeled by a different DPM, allowing us to account for measurement variations in a systematic manner.

For the sake of simplicity and to preserve conjugacy we assume that the covariance matrices of all local clusters associated with the global cluster are identical and limit the susceptibility

of local clusters to noise with their mean vectors. More specifically, $\mu_{jki} \sim G_{jk}$, $\sigma_{jki} = \sigma_k$, and F_{ϕ_k} is defined as

$$F_{\phi_k=\{\mu_k,\sum_k\}}=N\left(\mu_k,\frac{\sum_k}{\kappa_1}\right). \quad (3)$$

Note that the covariance matrix of the base distribution F_{ϕ_k} is a function of σ_k; hence conjugacy of the model is preserved. Conjugacy of the model is important since it enables us to implement a collapsed version of the Gibbs sampler. The scaling constant κ_1 adjusts the degree of dispersion of local means from the corresponding global mean. Posterior inference for the proposed model in (1) is performed by iteratively sampling local-cluster indicator variables for individual data points and global-cluster indicator variables for local clusters as described in detail in [20].

FC data matrices obtained from all available subjects in the training cohort (diseased and non-diseased samples) for a given tube were pooled together and the ASPIRE algorithm was run to identify local and global clusters (metaclusters) in the multi-dimensional point cloud defined by these data. Once global clusters and their local realizations were identified, the subject data for a given tube were characterized by the proportions of cells belonging to global clusters in the corresponding FC measurement, thus producing a vector of proportion values, i.e., a phenotype vector, for each tube. Since ASPIRE is a stochastic algorithm, the results obtained may differ slightly between runs. To account for this expected variability we run ASPIRE five times and report results averaged over these five runs.

D. Phenotypic Characterization for Test Samples

For the data and noise models discussed in Section II-C the posterior predictive distribution of a global cluster can be obtained in the form a multivariate student-t distribution with three parameters. In the following discussion a sample refers to a single FC mesaurement, i.e., a data matrix with rows representing cells and columns representing FC parameters, a global cluster refers to a cell type, and a data point refers to a cell.

$$p(\boldsymbol{x}_{jki}|D_{.k.})=stu-t(\hat{\boldsymbol{\mu}},\hat{\sum},v) \quad (4)$$

The location vector ($\hat{\mu}$), the scale matrix ($\hat{\sigma}$), and the degrees of freedom (v) are redefined for a global cluster as follows. Location vector:

$$\hat{\boldsymbol{\mu}}=\frac{\sum_{jkt:c_{jkt}=k}\frac{n_{jkt}\kappa_1}{(n_{jkt}+\kappa_1)}\overline{\boldsymbol{x}}_{jkt}+\kappa_0\mu_0}{\sum_{jkt:c_{jkt}=k}\frac{n_{jkt}\kappa_1}{(n_{jkt}+\kappa_1)}+\kappa_0} \quad (5)$$

Scale matrix:

$$\hat{\sum}=\frac{\sum_0+A_k}{\frac{\bar{\kappa}\,v}{\bar{\kappa}+1}} \quad (6)$$

Degrees of freedom:

$$v=m+\sum_{jkt:c_{jkt}=k}(n_{jkt}-1)-d+2 \quad (7)$$

where $\bar{\kappa}$ is defined as in (8)

$$\bar{\kappa}=\frac{(\sum_{jkt:c_{jkt}=k}\frac{n_{jkt}\kappa_1}{(n_{jkt}+\kappa_1)}+\kappa_0)\kappa_1}{\sum_{jkt:c_{jkt}=k}\frac{n_{jkt}\kappa_1}{(n_{jkt}+\kappa_1)}+\kappa_0+\kappa_1} \quad (8)$$

$D_{k\cdot}$ denotes the subset of cells sharing global cluster k across all samples, and $\bar{\boldsymbol{x}}_{jkt}$ and A_{jkt} are the sample mean and the scatter matrix for the local cluster t of global cluster k in sample j, respectively, n_{jkt} is the number of data points in the local cluster t of global cluster k in sample j, and A_k is the scatter matrix for the global cluster k. These statistics are defined as in (9).

$$\begin{aligned}\overline{\boldsymbol{x}}_{jkt} &= n_{jkt}^{-1}\sum_{jki:t_{jki}=t}\boldsymbol{x}_{jki}\\ A_{jkt} &= \sum_{jki:t_{jki}=t}(\boldsymbol{x}_{jki}-\overline{\boldsymbol{x}}_{jkt})\,(\boldsymbol{x}_{jki}-\overline{\boldsymbol{x}}_{jkt})^T\\ A_k &= \sum_{jkt:c_{jkt}=k}A_{jkt}\end{aligned} \quad (9)$$

After the distributions of global clusters are estimated during training, FC data matrices describing the test cases can be evaluated row-by-row using the posterior predictive distributions corresponding to global clusters. This results in a vector of posterior probability values with each value indicating the probability of a cell's belonging to one of the recovered global clusters (distinct cell immunophenotypes). The phenotypic characterization of a test case for a given tube at a given time point is achieved by averaging these probability vectors over all cells to get a phenotype vector for each FC measurement.

E. Classifier Training and Testing

The phenotype vectors corresponding non-diseased and AML cases in the training cohort were used to train a binary logistic regression classifier. Subsequently the classifier was applied to the test cases (diseased cases pre- and post-induction) in the test cohorts, and the probability of each test case's representing a phenotype similar to non-diseased was evaluated and recorded for all available time points. Additionally, leave-one-out probabilities for two hundred cases in the training cohort were also computed and recorded for use in

relative evaluation of test-case probabilities with respect to those of training cases. The LIBLINEAR package was used to train a L2-regularized logistic regression classifier [25].

III. Results

ASPIRE was previously evaluated for a one-class classification problem identifying AML cases as deviations from normal ones on the FlowCap II data set and extensive analysis comparing ASPIRE with several benchmark clustering techniques from the literature was reported with favorable results in [20].

In this study we target a more challenging clinical problem and further confirm the utility of ASPIRE in the clinical management of patients with AML by presenting results of a proof-of-concept study in predicting the direction of change in disease progression. Two separate cohorts of AML cases were used. The first contained data from 26 patients measured at two time points: pre- (t_0) and post-induction (t_1). The second contained data from 10 patients who relapsed after treatment. Three time points are accessible for this group: pre-induction, post-induction t_1, and post-induction t_2. The consensus diagnosis reached by manual flow cytometry analysis, cytogenetics, and histopatology evaluation was considered to be the ground truth for the purpose of our system validation.

A. Use of peripheral blood samples as a contrasting non-AML population

The ASPIRE-generated phenotype vectors corresponding to 100 non-AML and 100 AML cases were used to train a logistic regression classifier. Leave-one-out probabilities for these two hundred cases were also computed and represented for visualization. For the test cohorts, the probability of a patient's moving toward a non-AML phenotype at a given time point was evaluated by a logistic regression classifier using all available FC data for that patient (results from the seven-tube panel) processed by ASPIRE. A two-stage probabilistic evaluation was executed. In the first stage, evaluation was performed independently for each tube. In the second stage, phenotype vectors from all tubes were concatenated to form a higher-dimensional phenotype vector and evaluation was performed using these newly formed vectors.

For all cases in the first sub-cohort, when all tubes were considered, the increase in probability values from t_0 (pre-induction time point) to t_1 (post-induction time point) suggests that the proposed technique correctly identified the direction of disease progression for all twenty-six patients (Figure 1). When the same analysis was performed for individual tubes, results indicate that tubes 1, 2, 4, and 5 would be sufficient to correctly identify favorable disease progression for all patients tested (See Figures S1–S4 in the Supplement). Tubes 1 and 7 (Figure S5) yielded the largest change in the probability values from t_0 to t_1 (mean=0.44). This is still lower than the mean change of 0.55 obtained by using all tubes, which suggests that using all markers jointly may offer additional insight that would not be available when tubes are assessed independently.

For the second sub-cohort (10 patients who suffered relapse) with all tubes considered, the increase in probability values from t_0 to t_1 indicates that the proposed technique again correctly identified the favorable initial response for all ten patients (Figure 2). This result

matches the conclusion of the manual analysis of FC data performed by the trained flow cytometrist, as well as the results of cytogenetic and histopathology evaluations. However, when the probability values from t_1 to t_2 are considered, the decrease in probability values for nine out of ten cases indicates occurrence of relapse. This was a correct determination - confirmed by expert analysis - for all but one case. The mean increase in probability values from t_0 to t_1 is 0.53, whereas the mean decrease in probability values from t_1 to t_2 is 0.30.

When the same analysis was performed for individual tubes, it was noted that tube 1 alone again correctly identified the direction of change in disease progression for all ten cases from t_0 to t_1 as well from as t_1 to t_2 (Figure S6). For tube 1 the mean increase in probability values from t_0 to t_1 was 0.50, whereas the mean decrease from t_1 to t_2 was 0.29. These values are very similar to those obtained with all tubes combined, as well as higher than all other tubes analyzed individually.

Investigating further the single case for which the change in the direction of disease progression did not match the expert analysis performed by a hemopathologist when data from all tubes are used, we determined that the case indeed was a difficult case from a diagnostic perspective: manual cytometry data analysis and cytogenetics identified the case as a relapse, whereas the histopathological analysis diagnosed this case as “in remission.”

B. Use of stage I lymphoma bone-marrow samples as a contrasting non-AML population

The ASPIRE-generated phenotype vectors corresponding to 49 stage I lymphoma cases and 100 AML cases were used to train a logistic regression classifier. As before, leave-one-out probabilities for these 149 cases were computed and represented for visualization. For the test cohorts, the probability of a patient’s moving toward a non-diseased phenotype at a given time point was evaluated by a logistic regression classifier using all available FC data for that patient (results from the seven-tube panel) processed by ASPIRE. A two-stage probabilistic evaluation was executed. In the first stage, evaluation was performed independently for each tube. In the second stage, phenotype vectors from all tubes were concatenated to form a higher-dimensional phenotype vector and evaluation was performed using these newly formed vectors.

For all cases in the first sub-cohort, when all tubes were considered, the increase in probability values from t_0 (pre-induction time point) to t_1 (post-induction time point) suggests that the proposed technique correctly identified the direction of disease progression for all twenty-six patients (Figure 3). When the same analysis was performed for individual tubes, results indicate that tubes 2, 4, 5, 6, and 7 would be sufficient to correctly identify favorable disease progression for all patients tested (See the Suplemental Figures S7–S11). Tube 7 yielded the largest change in the probability values from t_0 to t_1 (mean=0.70). This is significantly higher than the mean change of 0.56 obtained by using all tubes as well as corresponding mean changes for the remaining six tubes for which the changes in the probability values from t_0 to t_1 are between 0.30 and 0.53.

For the second sub-cohort (10 patients who suffered relapse) with all tubes considered, the increase in probability values from t_0 to t_1 indicates that the proposed technique correctly identified the favorable initial response for all ten patients (Figure 4). This result matches the

conclusion of the manual analysis of FC data performed by the trained flow cytometrist, as well as the results of cytogenetic and histopathology evaluations. When the probability values from t_1 to t_2 are considered, the decrease in probability values for nine out ten cases indicates occurrence of relapse. This was a correct determination - confirmed by expert analysis - for all but one case. The mean increase in probability values from t_0 to t_1 is 0.53, whereas the mean decrease in probability values from t_1 to t_2 is 0.23.

When the same analysis was performed for individual tubes, it was noted that tube 5 alone correctly identified the direction of change in disease progression for all ten cases from t_0 to t_1 as well from as t_1 to t_2 (Figure S12). For tube 1 the mean increase in probability values from t_0 to t_1 was 0.36, whereas the mean decrease from t_1 to t_2 was 0.21. The mean increase in probability values from t_0 to t_1 is significantly lower than that obtained with all tubes combined (0.36 vs. 0.53). However, the mean decrease in probability values from t_1 to t_2 is comparable to that obtained with all tubes combined (0.21 vs. 0.23).

It is also interesting to note that the largest mean increase in probability values from t_0 to t_1 (0.61) as well as the largest mean decrease in probability values from t_1 to t_2 (0.25) is achieved for tube 7 (Figure S13). However, tube 7 predicts a decrease in probability values from t_0 to t_1 for one of the ten cases and an increase in probability values from t_1 to t_2 for two of the cases.

Upon further investigation we found that the single case for which ASPIRE failed to predict a decrease in probability from t_1 to t_2 when stage I lymphoma samples are used as normal cases is not the same case ASPIRE failed to predict a decrease in probability from t_1 to t_2 when blood samples were used as normal cases. However, for that specific case we have identified that the measurement at time t_2 has an artifact as it contains between only 500 to 1000 events across all seven tubes while all other measurements contain around 25,000 events.

IV. Discussion

Even though a number of previous reports demonstrated ability to automatically distinguish between normal and AML bone-marrow samples using computer-aided cytometry, the literature does not provide examples of autonomous algorithms in which different points in disease progression are associated with a meaningful probabilistic output that can be interpreted as reporting the direction of change (improvement or relapse). Our research targets this important area, using an innovative and original machine-learning methodology custom-developed for analysis of hematological FC results. The previously published binary statistical models did not generate a lot of excitement among clinical cytometrists, as the task of distinguishing pre-induction AML from normal samples is relatively easy, especially if abundant, well-characterized training samples are available. Automated classification becomes significantly more difficult when the system must account for disease diversity as well as for the inevitable presence of technical noise due to small differences in sample preparation, instrument operation, and data collection practices. The bar is additionally raised if access to the training samples is limited. All these obstacles were present in the processed data: the diseased training cohort represented samples spanning several years,

accounting for both biological diversity and technical noise, and the non-diseased set did not represented actual healthy bone-marrow samples, but a proxy sample of contrasting phenotype. The ASPIRE algorithm specifically developed to handle samples suffering from random effects performed well, extracting key biological features from the training data set and arriving at informative generalizations about the cell populations present.

In the cytometry sense, our system does not mimic a trained operator using preconceived and a priori-known information about informative cell populations and LAPs, but rather attempts to summarize the structure of the data point cloud by capturing the essential characteristics of the observed immunophenotypes. To accomplish this, in the first step the algorithm uses all the available training FC data (diseased bone-marrow samples and constrasting healthy peripheral-blood or stage I lymphoma samples) without any labels. This essential pre-processing part allows ASPIRE to generate a summary of relevant biology in the presence of inevitable experimental noise without using any labels and group information. In the second step we frame the problem as either anomaly detection or an “enhanced” detection of abnormality decrease.

The key limitation of using the ASPIRE system and similar Bayesian approaches is the non-deterministic output and computationally costly processing. Since ASPIRE relies on Markov-chain Monte-Carlo sampling techniques there could be slight variations between runs in terms of the number and size of global clusters discovered. To account for this variability, we run the algorithm multiple times and average results over multiple runs. This process raises the computational cost but fortunately is required only during the training phase, which is performed offline. The testing phase which is performed with cases blinded to the algorithm uses the global clusters discovered during training and involves assigning each cell in a FC sample to one of these global clusters. For a typical FC sample containing around 50,000 cells this deterministic classification can be performed with a run-time speed on the order of seconds and produces the same output in each repetition.

From the clinical FC perspective the important limitation of the demonstrated data-processing pathway is its incompatibility with established biology-driven data interpretation, in which the classification of a biological sample can be directly related to the judgment regarding proportion of cells in particular gates. The creation of these informative gates is not purely data-driven but guided by the previous experiences of an analyst interpreting the fluorescence-signal distribution in the space of selected biological parameters. Although ASPIRE formulates generalizations of biological subpopulations and represents them as metaclusters, the current version does not suggest which populations are “important” and cannot predict whether these statistically relevant populations are indeed biologically significant. From the practical perspective the need for direct visual interpretation of the automatically generated cytometry data analysis may be overestimated, as the goal of these processes is not to alleviate the burden of manual analysis in the first place. However, if such a need indeed arises one can imagine the implementation of an inverse step in which the covariates identified as important by the classifier are a subject to an inverse transform and represented at the end of the process as cluster boundaries overlaid onto traditional cytometry biaxial dot plots.

Regrettably, the relatively small cohort of patients used for the demonstration of the algorithm's capability does not allow us to investigate in depth its sensitivity and specificity across numerous important factors, such as AML subtype, patient age and/or overall health status, gene or chromosome changes in the cells belonging to abnormal populations, or the possibility that the studied leukemia occurred as a result of treatment for another different cancer. These are all important variables for which we hope to control in our further studies. It is also important to remember that the training phase did not use true normal bone-marrow samples. Again, at this phase of research it is impossible to determine if this leads to a biased classifier, which may perform differently for certain subgroups of possible AML immunophenotypes.

The ASPIRE algorithm is publicly available and can be tested by other FC researchers involved in studies of AML, as well as by scientists working on other hematological neoplasms.

V. Conclusions

This study provides the first example of a functional prototype of an automated non-parametric Bayesian clinical decision-support system that can not only recognize the difference between normal and abnormal samples, but most importantly also recognize the direction of change in disease progression on the basis of the FC bone-marrow data alone, without the need to supplement the data with morphology and/or genetic analysis. We believe that these results are an important first step for objective evaluation of disease status as well as for timely assessment of treatment strategies in the clinical management of AML patients.

Acknowledgments

This research was sponsored by the National Science Foundation (NSF) under Grant Number IIS-1252648 (CAREER), by the National Institute of Biomedical Imaging and Bio-engineering (NIBIB) under Grant Number 5R21EB015707, and in part by the National Cancer Institute (NCI) Cancer Center Support Grant 5P30 CA01605. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF and NIBIB.

References

1. [accessed: 2015-09-23] "Acute myeloid leukemia - surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) seer*stat database." http://seer.cancer.gov/statfacts/html/amyl.html
2. Shapiro, HM. Practical Flow Cytometry. 4. Hoboken, NJ, USA: Wiley-Liss; Mar. 2003
3. Wood BL, Arroz M, Barnett D, DiGiuseppe J, Greig B, Kussick SJ, Oldaker T, Shenkin M, Stone E, Wallace P. 2006 Bethesda International Consensus recommendations on the immunophenotypic analysis of hematolymphoid neoplasia by flow cytometry: optimal reagents and reporting for the flow cytometric diagnosis of hematopoietic neoplasia. Cytometry Part B, Clinical Cytometry. 2007; 72(Suppl 1):S1422.
4. Craig FE, Foon KA. Flow cytometric immunophenotyping for hematologic neoplasms. Blood. 2008; 111(8):3941–3967. [Online]. Available: http://www.bloodjournal.org/content/111/8/3941. [PubMed: 18198345]
5. Campo E, Swerdlow SH, Harris NL, Pileri S, Stein H, Jaffe ES. The 2008 who classification of lymphoid neoplasms and beyond: evolving concepts and practical applications. Blood. May.2011 117(19):50195032.

6. Vardiman JW, Thiele J, Arber DA, Brunning RD, Borowitz MJ, Porwit A, Harris NL, Beau MML, Hellstrm-Lindberg E, Tefferi A, et al. The 2008 revision of the world health organization (who) classification of myeloid neoplasms and acute leukemia: rationale and important changes. Blood. Jul.2009 114(5):937951.
7. Dhner H, Estey EH, Amadori S, Appelbaum FR, Bchner T, Burnett AK, Dombret H, Fenaux P, Grimwade D, Larson RA, et al. Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the european leukemianet. Blood. Jan.2010 115(3):453474.
8. Peters JM, Ansari MQ. Multiparameter flow cytometry in the diagnosis and management of acute leukemia. Archives of Pathology & Laboratory Medicine. 2011; 135(1):44–54. [Online]. Available: http://www.archivesofpathology.org/doi/full/10.1043/2010-0387-RAR.1. [PubMed: 21204710]
9. Inaba H, Coustan-Smith E, Cao X, Pounds SB, Shurtleff SA, Wang KY, Raimondi SC, Onciu M, Jacobsen J, Ribeiro RC, et al. Comparative analysis of different approaches to measure treatment response in acute myeloid leukemia. Journal of Clinical Oncology. 2012; 30(29):3625–3632. [PubMed: 22965955]
10. Campana D, Coustan-Smith E. Measurements of treatment response in childhood acute leukemia. The Korean journal of hematology. 2012; 47(4):245–254. [PubMed: 23320002]
11. Musen, MA., Middleton, B., Greenes, RA. Biomedical informatics. Springer; 2014. Clinical decision-support systems; p. 643-674.
12. Mansfield EA. Fda perspective on companion diagnostics: an evolving paradigm. Clinical Cancer Research. 2014; 20(6):1453–1457. [PubMed: 24634468]
13. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: the future just started. Cytometry Part A. Jul; 2010 77(7):705–713.
14. Robinson JP, Rajwa B, Patsekin V, Davisson VJ. Computational analysis of high-throughput flow cytometry data. Expert Opinion on Drug Discovery. Aug; 2012 7(8):679–693. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/22708834. [PubMed: 22708834]
15. Aghaeepour N, Finak G, Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH, et al. Consortium F, Consortium D. Critical assessment of automated flow cytometry data analysis techniques. Nature methods. 2013; 10(3):228–238. [PubMed: 23396282]
16. Bashashati A, Brinkman RR. A survey of flow cytometry data analysis methods. Advances in bioinformatics. 2009; 2009
17. Zare H, Bashashati A, Kridel R, Aghaeepour N, Haffari G, Connors JM, Gascoyne RD, Gupta A, Brinkman RR, Weng AP. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. American journal of clinical pathology. 2012; 137(1):75–85. [PubMed: 22180480]
18. Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences. 2014; 111(26):E2770–E2777.
19. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in flow cytometry data. Cytometry Part A. Jan; 2011 79(1):6–13. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21182178.
20. Dundar M, Akova F, Yerebakan HZ, Rajwa B. A non-parametric Bayesian model for joint cell clustering and cluster matching: Identification of anomalous sample phenotypes with random effects. BMC Bioinformatics. 2014; 15(1):314. [Online]. Available: http://www.biomedcentral.com/1471-2105/15/314/abstract. [PubMed: 25248977]
21. Yerebakan HZ, Rajwa B, Dundar M. The infinite mixture of infinite gaussian mixtures. Advances in Neural Information Processing Systems. 2014:28–36.
22. Tario, J., Wallace, P. Pathobiology of Human Disease. San Diego: Elsevier; 2014. Reagents and cell staining for immunophenotyping by flow cytometry; p. 3678-3701.
23. Ferguson TS. A Bayesian analysis of some nonparametric problems. Annals of Statistics. 1973; 1(2):209–230.
24. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. Journal of the American Statistical Association. 2001; 96(453):161–173. [Online]. Available: http://www.jstor.org/stable/2670356.

25. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. Liblinear: A library for large linear classification. The Journal of Machine Learning Research. 2008; 9:1871–1874.

Biographies

Bartek Rajwa is a Research Assistant Professor of Computational Biology in the Bindley Bioscience Center at Purdue University. Dr. Rajwa conducts research on biological pattern recognition and applications of statistical machine learning in cytomics and phenomics, with a special emphasis on high-throughput cytometry, high content imaging, and biological image analysis. Dr. Rajwas research focuses specifically on methods and techniques for quantitative description of cellular phenotypes using single cell-analysis approaches.

Paul K. Wallace is an internationally recognized expert in flow cytometry with a strong background in immunology and research interests in antigen processing and presentation. He is President Elect of the International Society for the Advancement of Cytometry (ISAC) and a Councilor of the International Clinical Cytometry Society (ICCS). In addition to serving as Director of the Flow and Image Cytometry at Roswell Park Cancer Institute (RPCI) in Buffalo, NY, he is a Professor of Oncology at Roswell Park and Laboratory Medicine at the University of Buffalo (SUNY).

Elizabeth A. Griffiths is an Assistant Professor in the Departments of Pharmacology and Therapeutics, Immunology and Medicine at Roswell Park Cancer Institute. She has extensive experience treating patients with Acute Myeloid Leukemia (AML) as an attending physician on the leukemia service where she treats more than 100 new patients with this diagnosis each year.

Murat Dundar is an Associate Professor of Computer and Information Science Department at IUPUI. His research expertise is in machine learning and data mining with a focus on non-parametric Bayesian models and inference, learning with partially-observed data, online and offline class discovery and modeling. His research is mainly driven by real-world problems in computer aided diagnosis/detection, hyper-spectral data analysis and remote sensing, bio-detection, flow cytometry data analysis, information retrieval, and topic modeling.

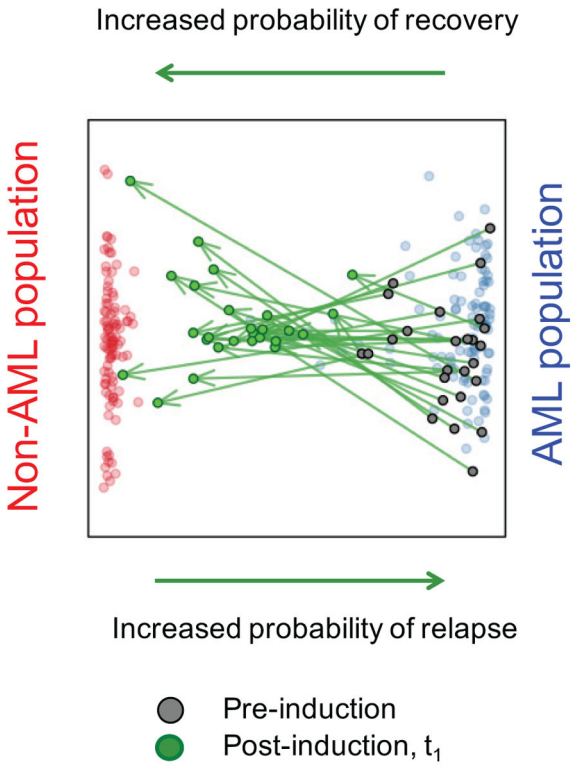

Fig. 1.

Probabilities obtained by the peripheral blood-based logistic regression classifier using phenotype vectors extracted by the ASPIRE algorithm from all the assay tubes available for the first cohort. The semi-transparent red points indicate non-diseased training samples, the semi-transparent green points denote AML training samples. The measures of individual test patients are shown as solid grey and green points. The arrows indicate the direction of change (disease progression) in the space defined by the training data. The results indicate that all the patients demonstrated positive change after therapy induction. The results match the known ground truth, as this patient sub-cohort responded well to therapy and subsequently recovered.

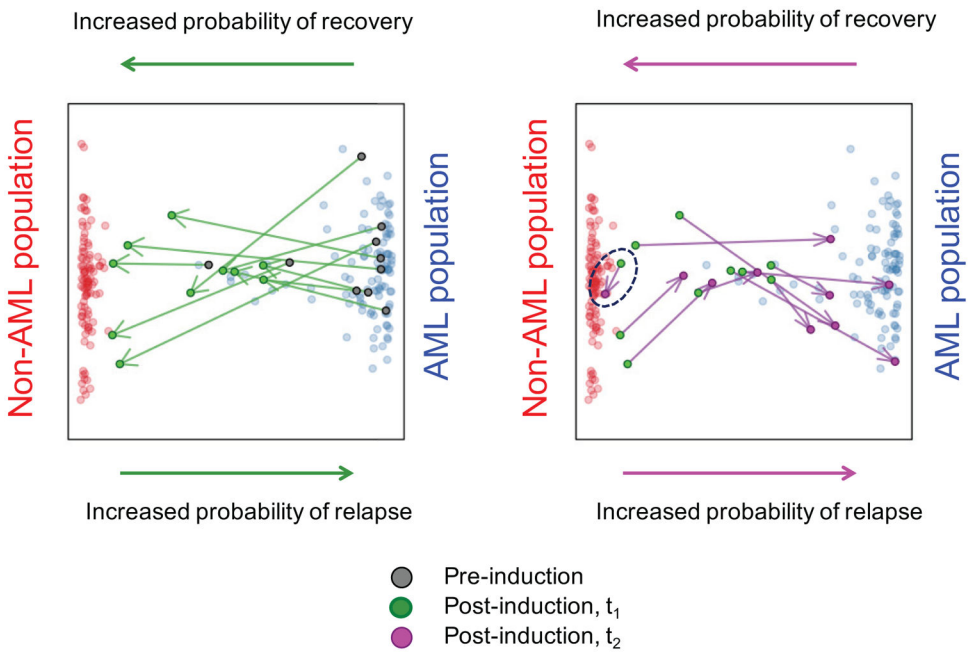

Fig. 2.

Probabilities obtained by the peripheral blood-based logistic regression classifier using phenotype vectors extracted by the ASPIRE algorithm from all the assay tubes for the second cohort. The measures of individual patients are shown as solid grey (pre-induction, t_0), green (post induction, t_1), and magenta points (post induction, t_2). The arrows indicating the direction of change in disease progression indicate that all but one patient demonstrated signs of relapse after initial positive response to the therapy (data point circled with a blue dotted line).

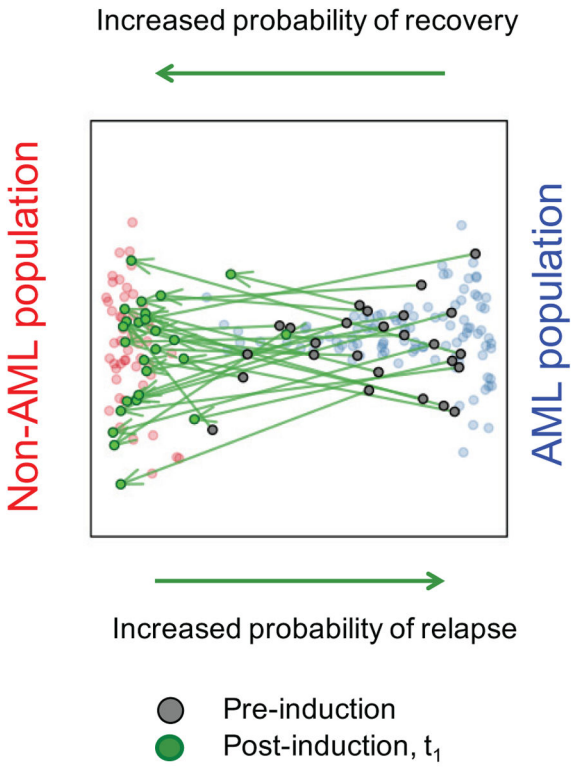

Fig. 3.

Probabilities obtained by the lymphoma-based logistic regression classifier using phenotype vectors extracted by the ASPIRE algorithm from all the assay tubes available for the first cohort.

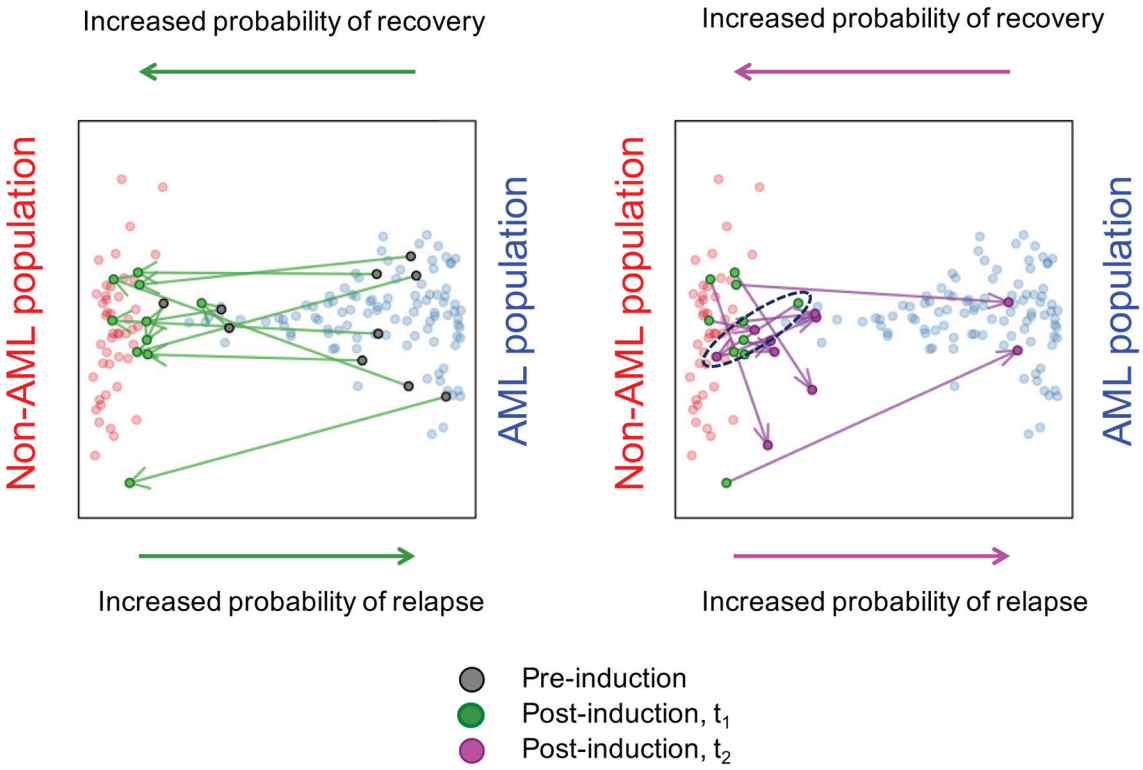

Fig. 4.

Probabilities obtained by the lymphoma-based logistic regression classifier using phenotype vectors extracted by the ASPIRE algorithm from all the assay tubes for the second cohort.

TABLE I

Tube IDs and corresponding marker and label combinations. Abbreviations: BD, BD Bioscience (San Jose, CA); BC, Beckman Coulter (Miami, FL); TF, Caltag Thermo/Fisher (Grand Island, NY); Dako (Carpinteria, CA)

	Marker combinations (mAb, source: clone)							
Tube ID	**FITC**		**PE**		**PcP/PECy5**		**APC**	
1	CD3	BD:SK7	CD14	BD:MØP9	HLADR PcP	BD:L243	CD45	BC:J33
2	CD11b	BC:Bear1	CD13	BD:L138	CD33 PECy5	BD:D3HL60.251	CD34	Dako: BIRMA-K3
3	CD15	BC:80H5	CD56	BC:NKH-1	CD7 PECy5	BC:8H8.1	CD34	Dako: BIRMA-K3
4	CD16	TF:3G8	CD32	TF:IV.3	CD45 PcP	BD:2D1	CD64	BD:10.1
5	CD38	BC:T16	CD10	BC:ALB1	CD19 PECy5	BC:J3.119	CD34	Dako: BIRMA-K3
6	CD41	TF:VIPL3	CD71	eBio:OKT9	CD45 PcP	BD:2D1	CD34	Dako: BIRMA-K3
7	CD57	BC:NC1	CD56	BC:NKH-1	CD33 PECy5	BC:D3HL60.251	CD3	TF:S4.1